DOCUMENT RESUME

ED 295 956 TM 011 717

AUTHOR Harker, Jill K.; Cope, Ronald T.
TITLE The Effect of Several Variables on Judgmentally-Obtained Cut Scores.
PUB DATE Apr 88
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; *Feedback; *Holistic Evaluation; *Interrater Reliability; *Licensing Examinations (Professions); Standards; Test Bias; *Test Theory
IDENTIFIERS *Angoff Methods; P Values

ABSTRACT

Cut scores obtained for licensure tests using different judgmental methods of standard setting (holistic, test blueprint, Angoff, and modified Angoff) were compared. Nineteen educators and practitioners participated in this study as judges. Pre- and post-test feedback (feedback of total- and low-group item p-value) ratings were obtained under the Angoff and modified Angoff methods. Results indicate: (1) significant differences between cut scores by modified versus unmodified methods for one of eight test forms, but no indication of relative upward or downward biasing for either method; (2) confirmation of the hypothesized tendency of judges to adjust ratings to be closer to feedback p-values; (3) small differences between cut scores derived from sets of first and second ratings; (4) moderately similar cut scores for three of the four tests across holistic, blueprint, and Angoff methods; and (5) very similar cut scores across methods for the remaining test. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED295956

The Effect of Several Variables on Judgmentally-Obtained Cut Scores

Jill K. Harker
Ronald T. Cope
The American College Testing Program

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

RONALD T. COPE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the
National Council on Measurement in Education
New Orleans, April, 1988

TM011717

The Effect of Several Variables on Judgmentally-Obtained Cut Scores

Abstract

This study compared cut scores obtained for licensure tests under different judgmental methods of standard setting: holistic, test blueprint, Angoff, and modified Angoff. Pre- and post-feedback (feedback of total- and low-group item p-values) ratings were obtained under the Angoff and modified Angoff methods.

Results show (1) significant differences ($p < .05$) between cut scores by modified versus unmodified methods for one of eight test forms, but no indication of relative upward or downward biasing for either method, (2) confirmation of hypothesized tendency of judges to adjust ratings to be closer to feedback p-values, (3) small differences (.01 to .03 on a proportion-correct scale) between cut scores derived from sets of first and second ratings, (4) for three of the four tests, moderately similar cut scores by holistic, blueprint, and Angoff methods; and for the remaining test, very similar cut scores by these methods.

Introduction and Rationale

One area in which absolute rather than normative examination standards are frequently required is professional licensure. Standards that are set grant successful examinees the opportunity to pursue the careers for which they have been trained. Although insight and support have been imparted to those of us working with professionals and lawmakers required to impose the false dichotomy on the continuum of competence (Shepard, 1980), the words "arbitrary" and "defensible" are presently being used in the same sentence with approximately the same frequency and discomfort that they were just after Glass (1978) and Popham (1978) discussed the nuances of their meanings with regard to standard setting. In the years that have elapsed, many of those struggling with the problems associated with standard setting have moved on to new, more interesting, and perhaps more fertile ground, extolling the virtues of item response theory (IRT) in various applications (van der Linden, 1982; Kane, 1987) in standard setting. These new directions give us hope for the future; however, many of us find licensing officials and their various boards having difficulty with much simpler concepts than those of IRT, and so we are still plugging away at trying to make the arbitrary decisions even more defensible while still using the older methods.

Many researchers have looked at different standard setting methods from different perspectives, comparing them, modifying them, praising them, or discrediting them (e.g., Brennan and Lockwood, 1980; Andrew and Hecht, 1976; Huynh and Saunders, 1979). Most of these articles begin with the assumption that some quantitative or quasi-quantitative procedure is better than none. One purpose of this study is to evaluate that assumption. Like competency, the cut score procedures presently in use can be placed on a continuum with holistic ratings (or guesses by experts) at one end and complex quantitative

or qualitative methodology at the other. Most of the procedures used lie somewhere between these two extremes.

Two procedures were chosen for comparison against holistic ratings given by judges participating in a licensure standard setting study. The holistic rating was the number that the judge felt in his or her "soul" to be the appropriate cut score to use. The two procedures chosen for comparison were (1) a somewhat more quantitative method suggested, in part, by Cangelosi (1984) in which judges rate test blueprints, and (2) the Angoff (1984) method (in two versions). The Angoff method was chosen because of its popularity, ease of use in the licensure setting, and apparent straightforwardness. Two versions of the Angoff method allowed us to investigate the nature of the differences that arise when judges are allowed to circle a given set of probabilities as opposed to writing in a number from 0.00 to 1.00 as an Angoff probability. Because groups of judges like these routinely request time-saving methods, we thought this investigation would be helpful.

Also, we investigated the effect of giving judges normative feedback in the form of item p-values for low-scoring and total examinee groups. We hypothesized that judges would adjust their initial, pre-feedback ratings to be closer to the p-values.

Method

A group of nineteen educators and practitioners chosen by a professional licensing board was convened to participate in the cut score study. This group was composed of five educators, one practitioner, and thirteen others who served both functions. These participants were sampled from different geographical areas of the particular jurisdiction and represented all types of practice in their profession. Since two levels of license designation exist (novice and advanced), two groups were created to independently rate the test

materials for each level of license. First, each of the educators was randomly assigned to one of the two groups, and then all other participants were randomly assigned to the same two groups. The group rating the novice licensee materials, Group 1, consisted of nine members (N = 9), and the group rating the advanced licensee materials, Group 2, consisted of ten members (N = 10).

All judges were asked to participate in three activities: (1) defining minimal competence by holistically choosing a cut score, (2) defining minimal competence by evaluating the test blueprints, and (3) defining minimal competence by rating test items. Each of these definitions of minimal competence yielded an estimated cut score. Each cut score estimate (averaged over judges) represented the judges' opinions of minimally acceptable performance from different perspectives. It was not assumed that any of these estimates would agree exactly. They served only as indicators to help in determining the final cut scores.

For the first activity, each judge was given a copy of the test blueprints for the examinations corresponding to the license level of his or her group and was asked to read it carefully. The judge was then asked to write down in a space provided his or her own best judgment of what the passing score should be for each novice or advanced licensee on each test. We have called these holistic ratings. These ratings were requested and done before the discussions of minimal competence took place and before judges had been given any kind of normative information or test items. All judges were aware of the cut scores currently in use for these examinations from publicly disseminated candidate information booklets.

For the second activity, judges were asked to rate each content area in terms of its frequency in practice and for its criticality to practice at the

entry level for the appropriate license. The judges were then asked to give a third rating of overall importance that would represent their personal, weighted averages of the frequency and criticality ratings. This was intended to help judges clarify to themselves the importance of each content area relative to candidate performance on the questions.

Finally, each judge was asked to determine the percentage of questions from each content area that a group of minimally-competent candidates should know. Because the test blueprint represented all of the essential areas of knowledge for licensure with the corresponding weights for each area, an estimated cut score for a content area could be determined if each judge decided the minimum number of items in each separate content area that should be answered correctly by minimally-competent candidates. This minimum number of items aggregated over all content areas would then yield an estimated passing score for the test.

The third activity of the study involved an Angoff and a modified Angoff rating procedure for two forms of each of the general examination (containing questions over the body of laws common to all jurisdictions) and the state-specific examinations for each group. Since the Angoff procedure has been modified many ways, definitions of our procedures using this method as a base follow.

1. Angoff Procedure: Judges were instructed to write their estimate of the probability that a minimally competent test taker will answer the item correctly. Their estimates were required to range from 0.00 to 1.00 inclusive. They were informed that the chance level is .25.
2. Modified Angoff Procedure: Judges were instructed to circle one of the following set of admissible probabilities to represent the

likelihood that a minimally competent test taker will answer the item correctly: .25 (chance level), .3, .4, .5, .6, .7, .8, .9, 1.00.

Notebooks with the materials to be rated were spiralled and passed out to group members. The judges rated in two stages. For the first rating, they supplied whichever Angoff probability was requested. After each judge finished the first rating on any one test form, the judge was given item analysis data from the most current administration of each item on that form (because this study was to evaluate the appropriateness of the current cut score, ample administration data were available). The data analysis provided sets of p-values based upon two examinee groups: (1) those with total test scores in the lowest 27% of examinees (low-group p-value), (2) total candidate population (total-group p-value). Judges were informed that they could change their first rating if they wished to do so but that it was unnecessary to do so if they felt their first rating was accurate. They were cautioned that the normative data should be considered supplementary information only and should not become the basis of their ratings. Throughout the rating process, the judges worked independently.

These activities resulted in four estimated passing scores for each test: the holistic rating, the blueprint rating, the first item rating, and the second item rating.

Results

Table 1 lists cut scores obtained under Angoff versus modified Angoff methods for both first and second ratings. Cut scores under these two Angoff methods were significantly ($p < .05$) different only for Form 1 of Test B. Note that, had the eight t-tests for either first or second ratings been independent of one another, one significant difference at the .05 level would

have been only slightly beyond chance expectation. Also, differences did not go in the same direction. For example, Test B, Form 1 cut scores under the Angoff method (written ratings) were .803 and .789 respectively, for first and second ratings; means by the modified Angoff method (circled ratings) were respectively .711 and .715 for first and second ratings. But for Form 2 of Test B, the Angoff method yielded lower cut scores, .752 and .765, than did modified Angoff, .812 and .817.

Insert Table 1 about here

Table 2, Table 3, and Figures 1, 2, 3, and 4 present information regarding influence of the item p-values given as statistical feedback upon item ratings averaged over judges. Comparison of mean first and second ratings in Table 2 shows relatively small changes between first and second ratings. Ratings of Test A, Form 1 and Test C, Form 1 showed the greatest

---

Insert Table 2 about here

---

differences, .03. The ratings tended to be adjusted to be more consistent with mean low- and total-group p-values. For example, the Test A, Form 1 first rating of .62 is lower than .65 and the total group p-value of .69. Here the adjustment was upward from .62 (first rating) to .65 (second rating). In contrast, the Test C, Form 1 first rating of .71, though close to the total group mean p-value of .73, exceeded the low group mean p-value of .59 by .14. Here the rating was adjusted downward to .68.

Table 3 presents intercorrelations of mean (over judges) first and second item ratings, low group p-values, and total group p-values. Low and total

group p-values showed approximately equal correlations with first ratings (respective mean correlations of .52 and .51) and with second ratings (.85 and .84).

Insert Table 3 about here

Figures 1-4 are overlay plots depicting changes of mean judge ratings of individual items for Test A, Form 2 (Figures 1 and 2) and Test C, Form 1 (Figures 3 and 4). Mean first ratings tended to be pulled downward for items of low p-value and upward for items of high p-value. These adjustments tended

---

Insert Figures 1-4 about here

---

to balance and to result in relatively small mean changes of rating but higher second rating/p-value correlations than first rating/p-value correlations (note slopes of regression lines).

Table 4 shows cut scores obtained under the blueprint and holistic methods and first and second ratings by the Angoff methods. The results read

---

Insert Table 4 about here

---

left to right in chronological order: judges first gave holistic ratings, then blueprint ratings, then Angoff ratings. Because judges gave blueprint and holistic ratings before seeing the tests, ratings by these methods could be obtained for tests but not for individual forms of the tests. Some judges misunderstood the directions for giving blueprint ratings and thereby produced

unusable data; thus blueprint ratings were based upon fewer judges than were the other ratings.

Test A cut scores ranged from .75 for the holistic method to .62 for the Form 1 first Angoff rating. This difference of .13 in the proportion-correct metric corresponds to 13 items on the 100-item Test A.

Test B cut scores agreed closely over all methods, the greatest difference being within the Angoff method: .75 for the Form 1 second rating versus .79 for the Form 2 second rating.

Test C cut scores under the blueprint and Angoff methods ranged from .68 to .71; the holistic method yielded a relatively high .76.

Test D cut scores by the Angoff method were all higher than the blueprint and holistic ones. Here the holistic result was the lowest, .73, whereas for Test A it was the highest at .75.

Table 5 illustrates the pass rates that would likely occur if the cut scores resulting from the various methods in this study were used. The cut scores are given as number of items correct and all scores were rounded up at 0.6. Because both the general and state specific examinations must be passed in order to obtain a license, the cut scores in the table are represented as composites of the number of items correct required on both tests. The test form cut scores resulting from the Angoff ratings were combined and averaged, because, in this program, the cut score is applied to new forms at each administration and the forms are not equated.

Insert Table 5 about here

The holistic ratings for the license designation allow for 53.6 percent of the examinees to pass, whereas the blueprint, the average first and average

second Angoff ratings are more closely in agreement, allowing from 69.4 to 73.2 percent to pass.

The holistic ratings for the advanced license designation allow for 55.6 percent of the examinees to pass, whereas the blueprint, the average first and average second Angoff ratings are, again, more closely in agreement, allowing from 32.2 to 46.6 percent of the examinees to pass.

Discussion

The significant differences between ratings by the Angoff (writing in ratings) and modified Angoff (circling one of a set of given ratings) methods are interesting. But interpreting these differences is difficult because over all tests and forms the differences are of different sign: one method does not yield consistently higher or lower results than the other. No clear indication of biasing by restricting the number or range of ratings--modified Angoff method--appears.

That the judges tended to adjust item ratings downward toward low p-value items and upward for high p-value items is not surprising. Neither is the tendency of these upward and downward adjustments to cancel each other to produce relatively slight overall changes of cut scores between first and second ratings. Of concern, however, is that in other studies statistical feedback has had quite different effects. A study by Cross, Impara, Frary, and Jaeger (1984) of mathematics and elementary education tests showed a decrease of standards after item p-value feedback under the Angoff method. A study by Cope (1987, April) of a different set of certification tests and judges from those of the present study showed consistent increases of standards. This apparent situation-specificity is cause for concern.

In the present study, all judges were given both low and total group p-values. The results show no indication that judges were more influenced by

one than the other. Next time we would like to try randomly assigning judges to receive only one of two or more levels of p-value feedback, such as those based upon lowest-scoring, marginal, or higher-scoring examinees, to see whether they produce different adjustments of standards.

Still, use of item p-values as normative feedback is hard to justify if one considers carefully what is being asked of judges. Their task seems appropriately modeled by item response theory (Cf, van der Linden, 1982; Kane, 1987). The Angoff method requires them to fix the minimum competence level θ^* of θ, the trait in question. Then they are required to estimate the probability $P_g(U_g = 1|\theta^*)$ that an examinee at level θ^* will answer a given item g correctly. $P(\theta^*)$ is not a p-value in the conventional sense. Yet by giving judges p-values we suggest that we want them to supply a p-value when in fact we are asking for something else. To judges conversant with item response theory we could more accurately instruct them as follows: "Decide the θ^* that represents the level of minimum competence, then look at the items and estimate their $P(\theta^*)$. You may first want to try estimating b, the difficulty parameter. But you will also need to consider the item discrimination, a. And because the item is multiple choice, you will also want to estimate c, the pseudo-guessing parameter, unless of course you are of the Rasch persuasion, in which case you can take one 'a' for this and all the other items and forget about 'c'." In practice, few if any judges know IRT. Yet this seems to be the sort of thought process expected of them. Their task, so explicated, is difficult. Instead of simply trying to model judge behavior by IRT, it is worth considering whether to instruct judges to think in this manner. Would doing so sharpen their judgment and thereby lead to more valid ratings and standards?

Nevertheless, the specious simplicity of p-values will likely lead to their continued use in practice as feedback. Accordingly, it behooves investigators to determine their effects under a range of circumstances.

The holistic methods of test blueprint ratings and of simply asking judges what they think the cut score ought to be show moderately close agreement with the Angoff method for three of the four tests and remarkably close agreement for Form B. Lacking any criteria other than possibly normative ones, it is worth considering at this point how to determine whether the time-consuming item-by-item judgmental methods of standard setting are really any better than the quick holistic ones.

It is noteworthy that, as information was provided to novice license category judges, the cut scores resulting from their ratings would allow greater numbers of examinees to pass, whereas the cut scores resulting from the advanced license category judges' ratings would allow greater numbers of examinees to fail.

REFERENCES

Andrew B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service. [Reprint of chapter in R. L. Thorndike (Ed.), Educational Measurement (2nd ed.) (pp. 508-600). Washington, DC : American Council on Education, 1971.]

Brennan, R. L. & Lockwood, R. E. (1981). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Cangelosi, J. S. (1984). Another answer to the cut-off score question. Educational Measurement: Issues and Practice, 3, 23-25.

Cope, R. T. (1987, April). A generalizability study of the Angoff method applied to setting cutoff scores of professional certification tests. Paper presented at the meeting of the American Educational Research Association, Washington, DC

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-129.

Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Kane, M. T. (1987). On the use of IRT models with judgmental standard setting procedures. Journal of Educational Measurement, 24, 333-345.

Popham, W. J. (1978). As always, provocative. Journal of Educational Measurement, 15, 297-300.

Shepard, L. A. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4, 447-467.

Van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 19, 295-308.

Table 1

Proportion-Correct Cut Scores by Angoff Versus Modified Angoff Methods for First and Second Ratings

		Rating 1				Rating 2			
Test	Form	Angoff	Modified Angoff	df	t	Angoff	Modified Angoff	df	t
A	1	.615	.617	7	-0.04	.653	.638	7	0.40
	2	.683	.696	7	-0.55	.684	.693	7	-0.35
B	1	.803	.711	8	2.54*	.789	.715	8	2.34*
	2	.752	.812	8	-2.10	.765	.817	8	-2.25
C	1	.733	.686	7	1.05	.706	.643	7	1.87
	2	.685	.695	7	-0.39	.670	.693	7	-0.85
D	1	.838	.768	8	1.64	.812	.751	8	1.32
	2	.801	.841	8	-1.21	.780	.824	8	-1.02

*$p < .05$

Table 2

Means and Standard Deviations of First and Second Item Ratings and Low- and Total-Group P-Values

Test	Form	Number of Items	Mean(SD) First Rating	Mean(SD) Second Rating	Mean(SD) Low-Group P-value	Mean(SD) Total-Group P-value
A	1	100	.62 (.07)	.65 (.09)	.65 (.18)	.69 (.14)
	2	100	.69 (.06)	.69 (.08)	.65 (.18)	.79 (.13)
B	1	100	.76 (.07)	.75 (.10)	.63 (.24)	.77 (.17)
	2	100	.78 (.08)	.79 (.09)	.71 (.22)	.82 (.14)
C	1	40	.71 (.06)	.68 (.09)	.59 (.18)	.73 (.15)
	2	40	.69 (.06)	.68 (.10)	.61 (.20)	.75 (.17)
D	1	50	.80 (.07)	.78 (.11)	.62 (.25)	.73 (.21)
	2	50	.82 (.06)	.80 (.08)	.65 (.27)	.76 (.19)

Table 3

Correlations of Mean First Rating, Mean Second Rating Correlation and Feedback p-Values

Judge Group	Test	Form	Number of Items	First Rating/ Second Rating	First Rating/ Total Group p-value	Second Rating/ Total Group p-value	First Rating/ Low Group p-value	Second Rating/ Low Group p-value	Total Group p-Value/Low Group p-value
1	A	1	100	.84	.42	.78	.42	.79	.95
		2	100	.91	.66	.86	.64	.85	.95
2	B	1	100	.77	.41	.85	.34	.81	.90
		2	100	.87	.54	.83	.54	.79	.86
1	C	1	40	.80	.50	.88	.46	.86	.96
		2	40	.86	.65	.92	.66	.93	.95
2	D	1	50	.90	.63	.83	.65	.85	.91
		2	50	.77	.37	.84	.34	.81	.93
	$\bar{X}$			.84	.52	.85	.51	.84	.93
	s			.06	.12	.04	.14	.05	.03

Table 4

Multimethod Estimates of Cut Scores

Judge Group	Test	Holistic N	Holistic Mean	Holistic SD	Blueprint N	Blueprint Mean	Blueprint SD	Form	First Angoff Rating N	First Angoff Rating Mean	First Angoff Rating SD	Second Angoff Rating Mean	Second Angoff Rating SD
1	A	9	.75	.03	5	.67	.13	1	9	.62	.07	.65	.05
								2	9	.69	.04	.69	.04
2	B	10	.76	.03	9	.76	.05	1	10	.76	.07	.75	.06
								2	10	.78	.05	.79	.05
1	C	9	.76	.06	5	.68	.19	1	9	.71	.07	.68	.06
								2	9	.69	.04	.68	.04
2	D	10	.73	.04	9	.76	.04	1	10	.80	.07	.78	.08
								2	10	.82	.07	.80	.07

Note. N is number of judges whose cut scores were averaged; results are expressed as proportions of number of test items.

Table 5

Passing Score and Corresponding Estimated Pass Rates by Method

Judge Group	Combined Test Forms	Holistic Cut Score & Pass Rate	Blueprint Cut Score & Pass Rate	Comb. Test Forms	First Angoff Cut Score & Pass Rate	Second Angoff Cut Score & Pass Rate
1	A/C	75/30 53.6	67/27 73.2	A_1C_1	62/28 69.4	65.27 75.1
				A_1C_2	same	same
				A_2C_1	69/28 66.2	69/27 71.5
				A_2C_2	same	same
			Mean		64/28 69.4	67/27 73.2
			Approximate (SD)		(3.5)/(0.0)	(2.0)/(0.5)
2	B/D	76/36 55.6	76/38 46.6	B_1D_1	76/40 32.3	75/39 42.3
				B_1D_2	76/41 23.3	75/40 32.3
				B_2D_2	78/40 31.7	79/39 40.2
				B_2D_2	78/41 22.8	79/40 31.7
			Mean		77/40 32.2	77/39 41.8
			Approximate (SD)		(1.0)/(0.5)	(2.0)/(0.5)

Notes. In this licensure program, the general and state specific examinations must both be attempted in order for a candidate for obtain a license.

Pass rates reflect the percentage of candidates passing having achieved the number right corresponding to the cut scores given in the table. Pass rates are based on N = 4321 for Judge Group 1 (novice license) and N = 189 for Judge Group 2 (advanced license).

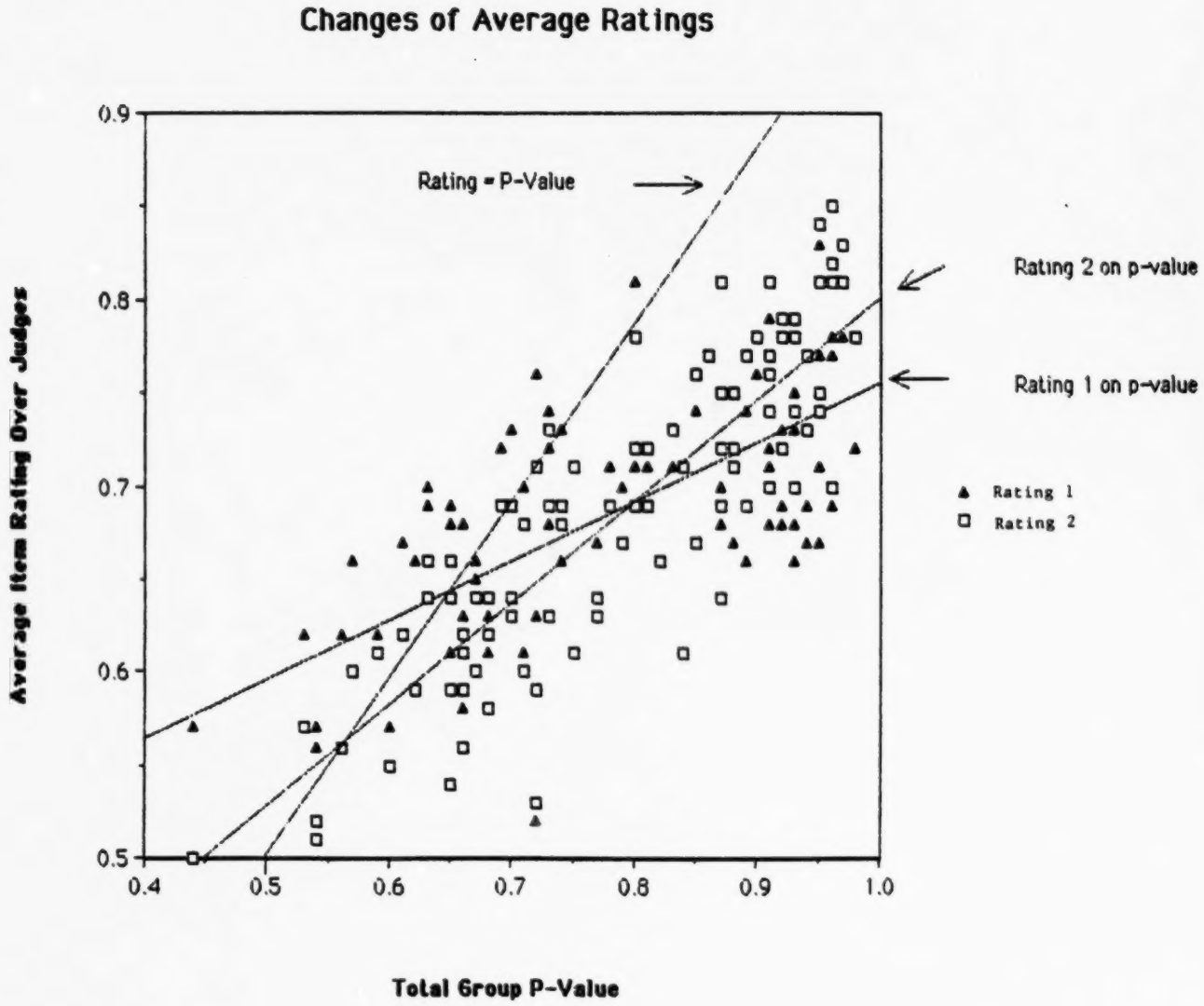

Figure 1. Overlay plot of average first and second item ratings against item p-values for total examinee group, Form 2 of Test A.

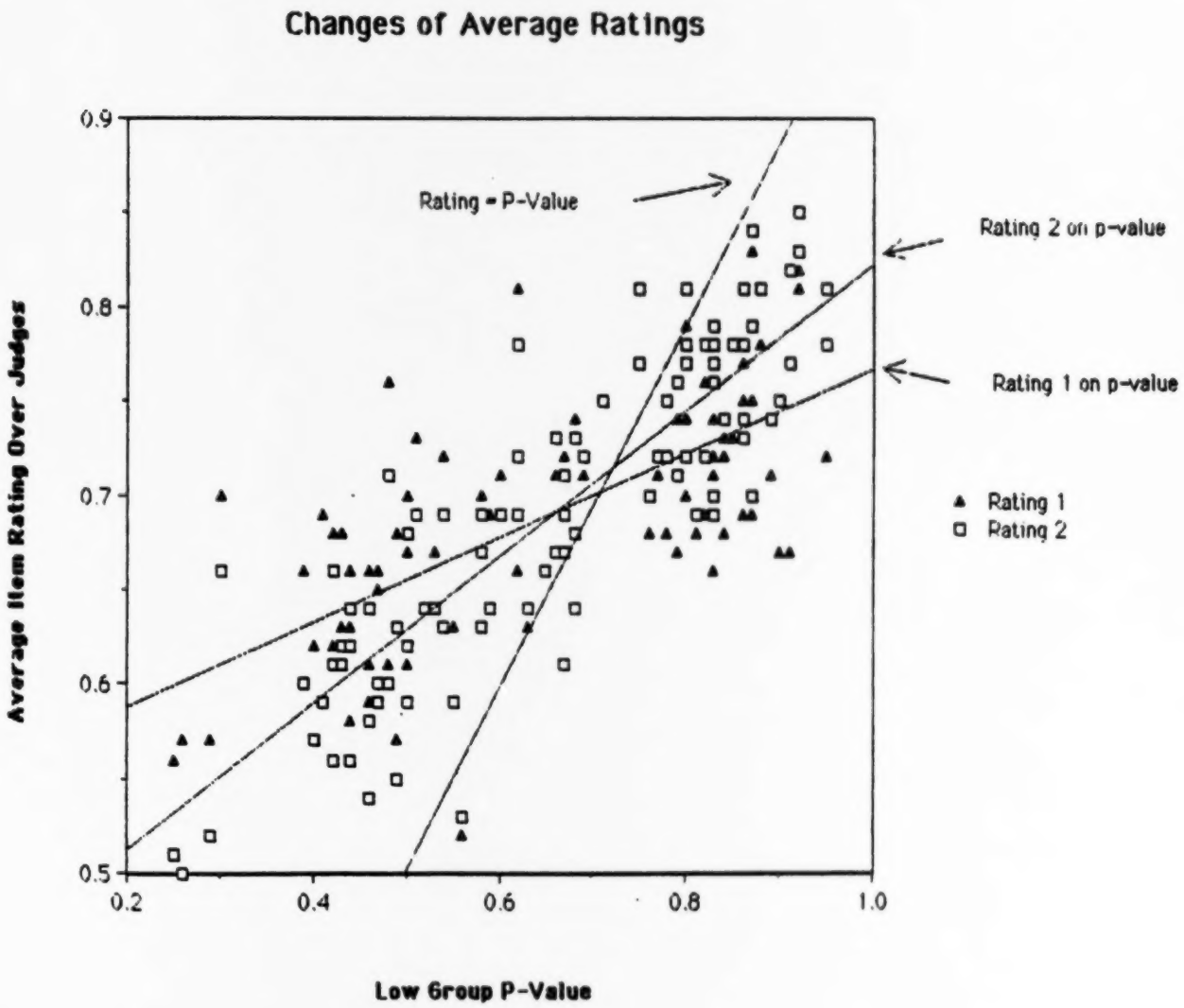

Figure 2. Overlay plot of average first and second item ratings against item p-values for low-scoring examinee group, Form 2 of Test A.

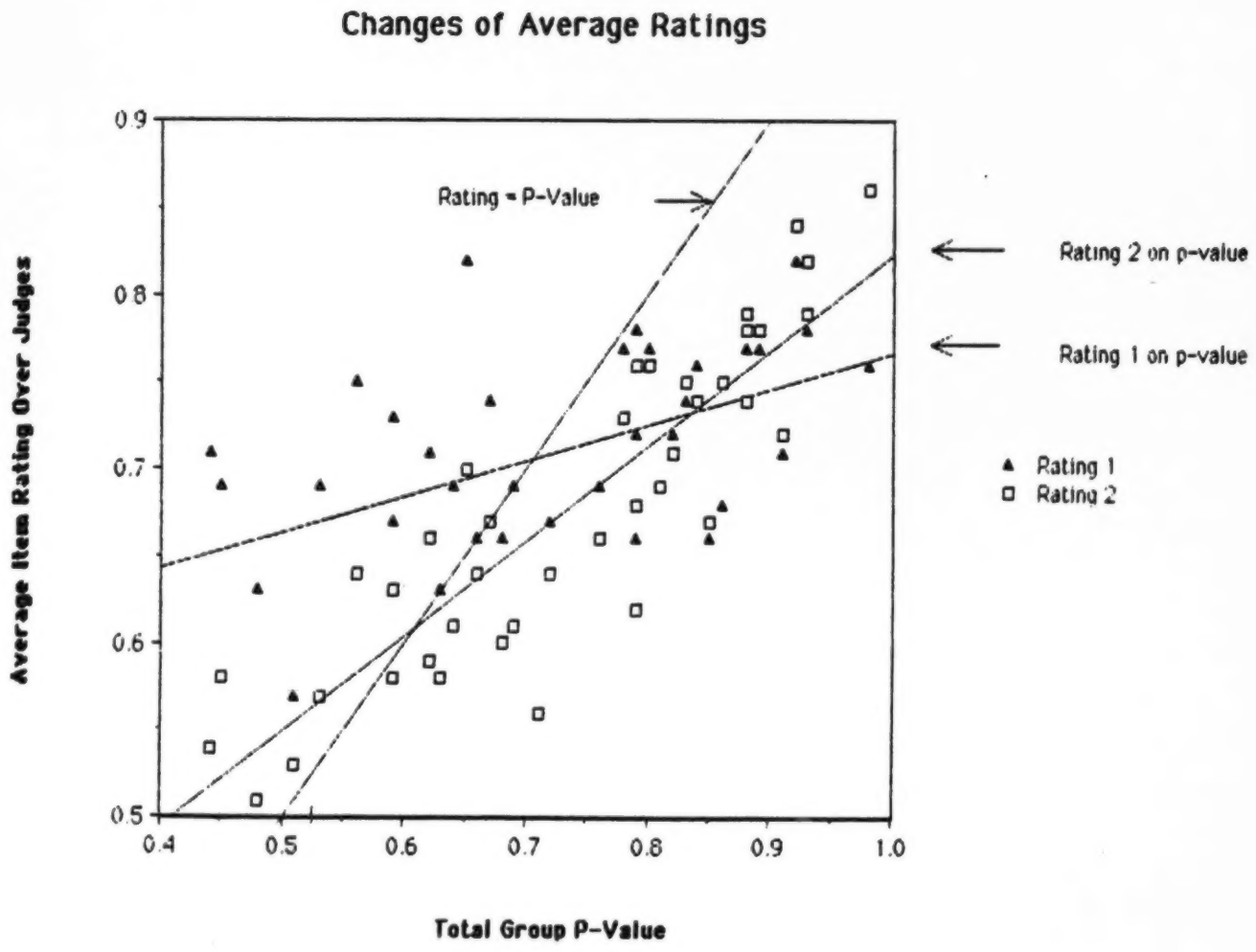

Figure 3. Overlay plot of average first and second item ratings against item p-values for total examinee group, Form 1 of Test C.

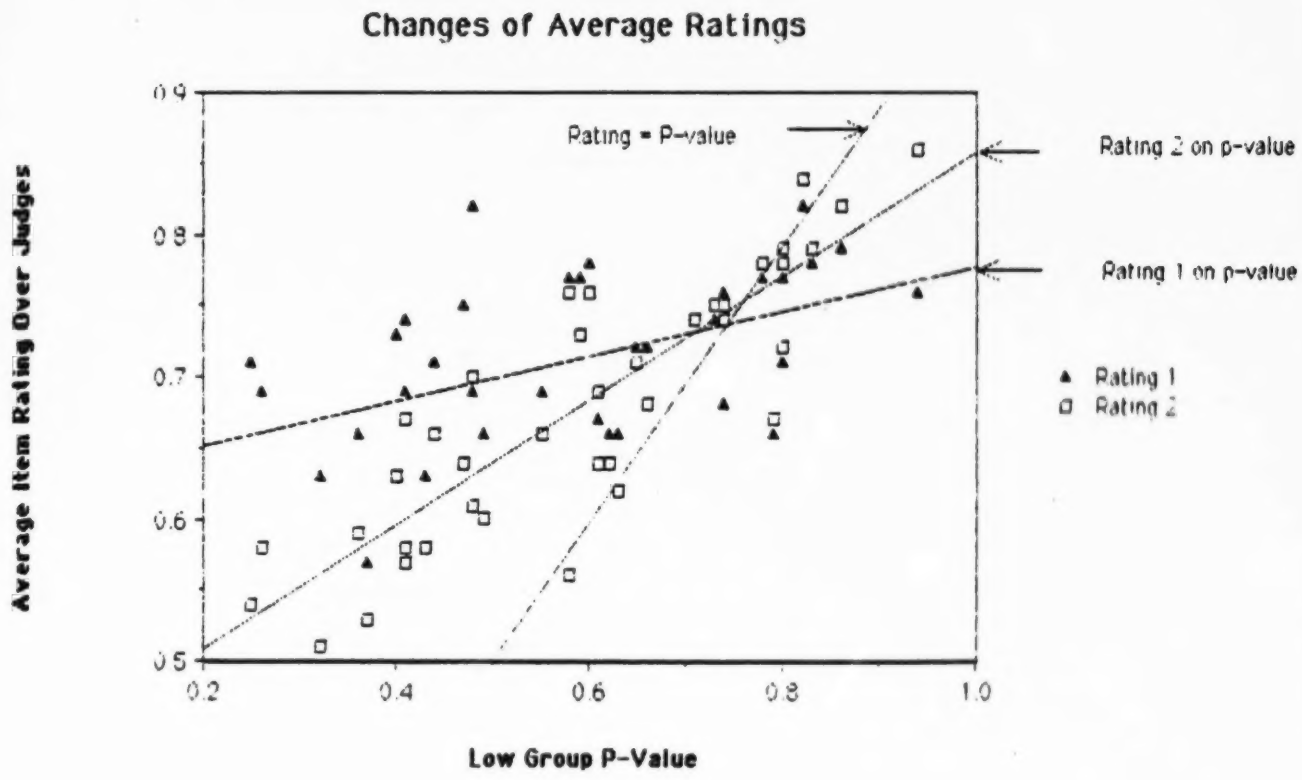

Figure 4. Overlay plot of average first and second item ratings against item p-values for low-scoring examinee group, Form 1 of Test C.